The Meaning of Taflig: Distributional Similarity for Rare Words

Julie Weeds

juliewe@cogs.susx.ac.uk

School of Cognitive and Computing Sciences

1 Introduction

My area of research is the relationship between a word's semantics (or meaning) and its syntactic behaviour. The underlying idea is that if two words mean similar things then they will occur in documents in similar contexts. In other words, in a given sentence it is generally possible to substitute each word with a synonym (i.e. another word with the same meaning). For example, in the following sentences we have the unknown word '*taflig*'.

```
I like TAFLIG.
TAFLIG costs more than it used to.
I drink TAFLIG.
Drinking TAFLIG gives me a hangover.
There are lots of TAFLIG cans in
  the recycling bin.
```

By considering what words could possibly occur in each sentence instead of *taflig*, we can determine that *taflig* means something similar to *beer*.

2 Related Work

Various researchers (e.g. Lin (1998) and Grefenstette (1994)) have used this idea to automatically construct thesauruses. They do this by analysing large amounts of text and extracting for each word a description which consists of a set of dependency tuples. The exact form of a dependency tuple varies from researcher to researcher, but it might, say, consist of the word, another word with which it cooccurs and the syntactic relationship between the two words. Associated with each tuple is also the frequency of occurrence of the tuple in the text. For example, the description extracted of *taflig* from the above sentences might look something like this:

```
(taflig,like,obj-of) - 1
(taflig,cost,subj-of) - 1
(taflig,drink,obj-of) - 2
(taflig,can,noun-mod-of) - 1
```

Once a description has been constructed for each word, the similarity between any two words can be computed using one of a variety of similarity measures.

This simple technique leads to fairly good results and most research to date has focussed on the similarity measure used and the dependency relationships considered. However, there are a number of other issues, two of which I will briefly discuss here.

3 Polysemy

Firstly, many words in English are polysemous i.e. they have more than one sense. For example, the word *bank* has a number of senses which include a *financial institution* sense and a *river bank* sense. Our hypothesis is that it is the meaning of a word which leads to syntactic similarity between words. Polysemy is therefore a problem because this syntactic similarity will exist between word senses not between word forms. Unless the sentences examined are sense tagged (i.e. each word is labelled with the sense being used), which is very rare, the description of a word conflates all of the senses. This can lead to a down-grading of all the similarity scores since a word similar to one sense of another word is unlikely to be similar to all of its senses. However, generally, unless one sense is particularly dominant (i.e. by far the most frequent) a mixture of similar words for each sense is found. It would be useful to be able to automatically determine to which sense each 'similar' word was similar.

4 Reliability of Similarity Scores

The second problem, which makes the first one harder to solve, is that we do not know how reliable our similarity scores are. This reliability is affected by a number of factors including the number of tuples we have seen for each word, the information content of each tuple (for example "I drink

taflig" has a higher information content than "I like taflig" because drink is a more strongly selecting verb than like) and the number of different syntactic contexts/tuple types a particular word can occur in.

5 A New Approach

In order to tackle these problems, I consider the similarity data as a fully-connected weighted graph. In this graph, nodes represent words and the weight on each link is the similarity score between the two words it connects. On first sight it might seem that we can easily identify separate senses of a word. If a word has two senses then we might expect a pattern such as that exhibited by the simplified graph in Figure 1 to arise.

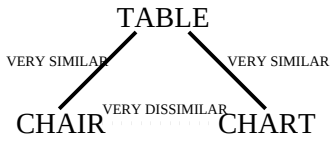

Figure 1: Multiple senses of the word *table*

However, this is not the only situation which will give rise to such a pattern. When our similarity scores are unreliable, as illustrated in Figure 2, this pattern may arise by chance. In this example, if we were to collect more data (i.e. more sentences), we would probably find that *cat* and *rabbit* would become more similar. However, as we consider rarer words it becomes increasingly difficult to collect more data. Hence, can we determine when two words should be more similar without collecting more data?

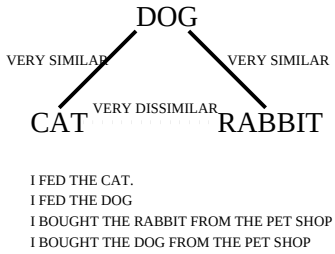

Figure 2: A single sense of the word *dog*

My research looks at how to assign a reliability or confidence score to each link and whether the problem discussed above can be resolved when we consider patterns between larger numbers of nodes. For example, Figure 3 shows the patterns we might expect to see arise in the above cases when we look at five nodes.

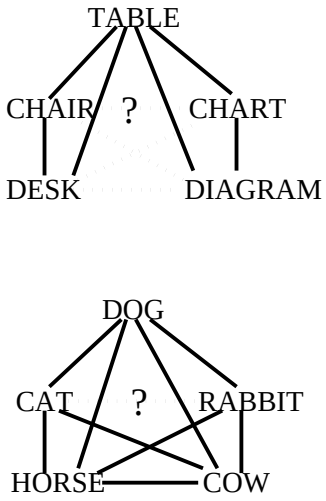

Figure 3: Looking at more nodes

References

Grefenstette, G. (1994). *Explorations in automatatic thesaurus discovery.* Dordrecht: Kluwer.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING-ACL*. Montreal.

Biographical Information Julie Weeds graduated from Trinity Hall, Cambridge in 1998 with a 1st in Computer Science. She then spent a year in London training as a patent attorney. During this time she realised that she didn't like the real world very much and so fled back to Cambridge in 1999 to do an MPhil in Computer Speech and Language Processing. In October 2000, she arrived at COGS to begin her DPhil with the Natural Language Processing Group under the supervision of David Weir.
